

# Combinatorics of $k$ -Interval Cospeciation for Cophylogeny

Jane Ivy Coons, Joseph Rusinko

## Abstract

We show that the cophylogenetic distance,  $k$ -interval cospeciation, is distinct from other metrics and accounts for global congruence between locally incongruent trees. The growth of the neighborhood of trees which satisfy the largest possible  $k$ -interval cospeciation with a given tree indicates that  $k$ -interval cospeciation is useful for analyzing simulated data.

## Index Terms

Cophylogenetics,  $k$ -interval cospeciation, tree metrics, combinatorics



## 1 INTRODUCTION

THE field of phylogenetics aims to discern evolutionary relationships between species using biological data and mathematics. A relatively new subfield of phylogenetics, known as *cophylogenetics*, concerns the evolutionary processes of groups of taxa, or taxonomical units of any rank, that we believe are evolving concomitantly. Cophylogenetics is applicable, for instance, in the cases of host-parasite coevolution and the coevolution of a species and the genes within that species [1], [2].

In cophylogenetics, we require tools such as reconstruction algorithms and distance metrics to determine the evolutionary processes of the taxa we are studying and how these processes are related. There are many reasons why host-parasite and gene-species trees do not match exactly, such as a parasite's failure to speciate in reaction to a host speciation, parasite extinction, or independent speciation of the parasite [1]; however, there are expected similarities between the trees. Therefore,  $k$ -interval cospeciation is defined in order to measure these differences between trees, and quantify their level of congruence [1].

This is applicable to cophylogeny when, for instance,  $T_1$  represents a host tree,  $T_2$  represents a parasite tree, and a pair of taxa,  $A$  and  $a$  are associated if  $a$  is a parasite that lives on a host,  $A$ . Two  $n$ -taxa trees,  $T_1$  and  $T_2$ , satisfy  $k$ -interval cospeciation if for all pairs of corresponding taxa on the trees, the number of edges between two taxa,  $A$  and  $B$ , on  $T_1$  is within  $k$  of the number of edges between their corresponding taxa,  $a$  and  $b$ , on  $T_2$ . We can also apply this to two possible trees on the same set of taxa – such as several gene trees – where two taxa on different trees are associated if they belong to the same species. In [1], the authors pose an open problem to characterize the combinatorial properties of  $k$ -interval cospeciation; this paper is a response to that open problem.

Several studies of host-parasite relationships and of gene-species trees compare cophylogenetic trees in order to understand coevolution [2], [3], [4] [5]. In [3], the authors constructed phylogenetic trees for Caryophyllaceae and Microbotryum (hosts and parasites, respectively) using a Bayesian 50% majority rule consensus tree approach. They noted that on a large scale, the trees appear quite congruent, though they have many local incongruences. They were able to determine this through visual inspection, as their trees were relatively small; however, for larger trees, this may prove difficult. Our research indicates that  $k$ -interval cospeciation is a viable means of quantifying the degree of this global congruence. For instance, the host-parasite trees in Figure 4 of [2], which describe the evolutionary processes of pelicaniform birds and the lice that live on them, satisfy 5-interval cospeciation. This relatively low  $k$ -interval cospeciation accounts for the high level of global congruence observed between these trees.

Distance metrics such as  $k$ -interval cospeciation are also useful for analyzing how phylogenetic methods perform on simulated data; however, some metrics are more suited to this task than others because of their combinatorial properties. We show that, because of the growth of the neighborhood of trees that satisfy the largest possible  $k$ -interval cospeciation with a given tree,  $k$ -interval cospeciation is more useful than other common distance metrics for analyzing data generated from simulations.

- J. Coons is a student at the State University of New York at Geneseo and a participant in the Winthrop University Research Experience for Undergraduates. E-mail: janeivycoons@gmail.com
- J. Rusinko is with Winthrop University. Email: rusinkoj@winthrop.edu

## 2 BACKGROUND

An *unrooted binary tree* is a phylogenetic tree where each non-leaf vertex has degree three. At each leaf of the tree is a *taxon* (plural: taxa), such as a gene, species or population. We define a *cherry* to be a pair of taxa with exactly two edges between them, and a *caterpillar tree* to be a phylogenetic tree with exactly two cherries. Since  $k$ -interval cospeciation is a discrete distance metric, we do not consider branch lengths.

**Definition 1.** Let  $\varepsilon_H(A, B)$  be the number of edges in the shortest path between taxa  $A$  and  $B$  on  $T_H$  and  $\varepsilon_P(a, b)$  be the number of edges in the shortest path between taxa  $a$  and  $b$  on  $T_P$ . Then  $T_H$  and  $T_P$  satisfy  $k$ -interval cospeciation ( $k$ -IC) if the following inequality holds for all  $A$  and  $B$ , and corresponding  $a$  and  $b$ :

$$|\varepsilon_H(A, B) - \varepsilon_P(a, b)| \leq k. \quad (1)$$

Note that, by (1), if two trees satisfy  $k$ -IC, they also satisfy  $(k + x)$ -IC for any positive integer,  $x$ . So, we say that two trees satisfy *precise  $k$ -IC* if they satisfy  $k$ -IC and they do not satisfy  $(k - 1)$ -IC. If  $T_1$  and  $T_2$  satisfy precise  $k$ -IC, then  $d(T_1, T_2) = k$ .

**Theorem 1.** *Precise  $k$ -IC is a tree metric.*

*Proof:* For any tree,  $d(T, T) = 0$ , since the number edges between each pair of taxa must remain the same. Additionally, for any two  $n$ -taxa trees,  $T_1$  and  $T_2$ ,  $d(T_1, T_2) = d(T_2, T_1)$ . So, in order for precise  $k$ -IC to be a distance, we must show that it satisfies the triangle inequality.

Let  $T_1, T_2, T_3$  be  $n$ -taxa trees for which  $d(T_1, T_2) = x$ ,  $d(T_2, T_3) = y$ , and  $d(T_1, T_3) = z$  for some  $x, y, z \in \mathbb{Z}_{\geq 0}$ . We must show that  $z \leq x + y$ . Consider arbitrary leaves,  $i_1, j_1 \in T_1$ ,  $i_2, j_2 \in T_2$  and  $i_3, j_3 \in T_3$ . By definition of  $k$ -IC,

$$-x \leq \varepsilon_1(i_1, j_1) - \varepsilon_2(i_2, j_2) \leq x, \quad (2)$$

$$-y \leq \varepsilon_3(i_3, j_3) - \varepsilon_2(i_2, j_2) \leq y, \quad (3)$$

$$-z \leq \varepsilon_1(i_1, j_1) - \varepsilon_3(i_3, j_3) \leq z. \quad (4)$$

Adding 2 and 3 yields  $-x - y \leq (\varepsilon_1(i_1, j_1) - \varepsilon_2(i_2, j_2)) - (\varepsilon_2(i_2, j_2) - \varepsilon_3(i_3, j_3)) \leq x + y$ , which simplifies to  $|\varepsilon_1(i_1, j_1) - \varepsilon_3(i_3, j_3)| \leq x + y$ . So, by definition of  $k$ -IC,  $T_1$  and  $T_3$  satisfy  $(x + y)$ -IC. So, in order for  $T_1$  and  $T_3$  to precisely satisfy  $z$ -IC,  $z \leq x + y$ , as needed. Therefore, precise  $k$ -IC satisfies the triangle inequality. So, precise  $k$ -IC is a distance.  $\square$

## 3 PRECISE $k$ -INTERVAL COSPECIATION AND OTHER DISTANCES

A comparison of precise  $k$ -interval cospeciation to other common discrete tree distances indicates that  $k$ -IC is a novel metric. We say that a distance metric,  $d_a$ , *determines* another distance,  $d_b$  if there exists a bijection,  $f$  such that  $f(d_a(T_1, T_2)) = d_b(T_1, T_2)$  for all  $T_1$  and  $T_2$ . We found that several of the most commonly used discrete distance metrics do not determine  $k$ -IC, and that  $k$ -IC reflects global similarities between trees which other distance metrics do not. This indicates that  $k$ -IC is a unique metric with properties that are not represented by current phylogenetic distances.

One extremely common distance metric is the nearest neighbor interchange distance. A *nearest neighbor interchange* (NNI) is a tree rearrangement operation in which we switch two subtrees of a tree that are joined by a single edge. So, the nearest neighbor interchange distance between  $T_1$  and  $T_2$  is the smallest number of NNIs it takes to transform  $T_1$  into  $T_2$ , or vice versa. Finding the NNI distance between two given trees is an NP-complete problem [6].

**Theorem 2.** *Determining the  $k$ -IC between two trees is computable in polynomial time.*

*Proof:* This is evident from the fact that the shortest path between two nodes of a tree can be computed at  $O(n \log n)$  [7], and from the fact that we must compute these path  $\binom{n}{2}$  times. So, computing  $k$ -IC can be done at  $O(n^3 \log n)$ .  $\square$

For this reason, we hoped to find a relationship between the NNI distance and  $k$ -IC in order to bound the NNI distance more efficiently [1].

Huggins et al. proved that two trees satisfy 1-IC if and only if their NNI distance is one [1]. However, they also introduced a counterexample to the possibility of  $k$ -IC providing an upper bound for the NNI distance for any  $k > 1$ . These counterexample trees,  $C_{3x}$  and  $D_{3x}$ , can be built by adding rooted triples on the same leaf set, but with differing cherries to the pendant edges of two  $x$ -taxa caterpillar trees,  $C$  and  $D$ . Fig. 1a and Fig. 1b are examples of these trees on 15 taxa.

Since no NNI operation is able to alter two of these rooted triples at the same time, the NNI distance between  $C_{3x}$  and  $D_{3x}$  is  $x$ . However, the two trees satisfy 2-IC. So,  $k$ -IC does not provide an upper bound for the NNI distance between trees. Therefore, the NNI distance does not determine  $k$ -IC. However,  $k$ -IC does provide a strict lower bound on the NNI distance.

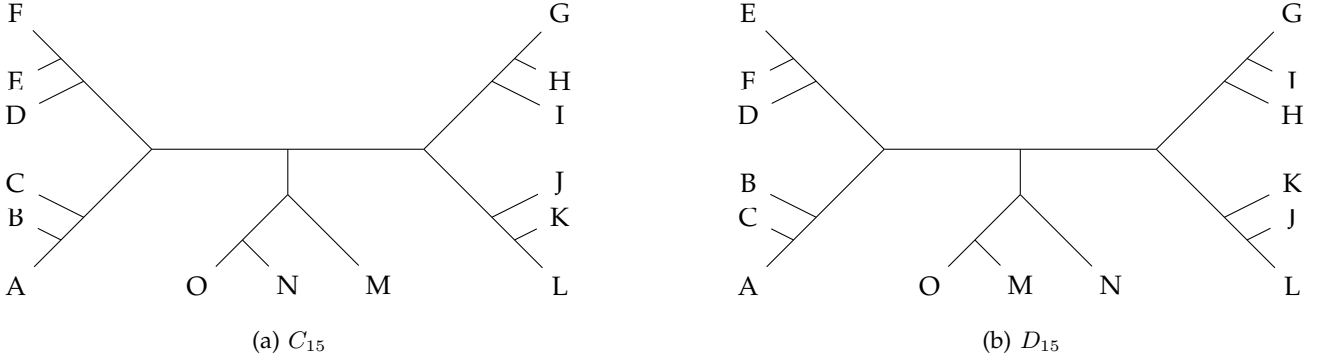


Fig. 1: Counterexample Trees Introduced in [1].



Fig. 2

**Theorem 3.** Suppose two phylogenetic trees,  $T_1$  and  $T_2$  precisely satisfy  $k$ -IC. Then  $d_{NNI}(T_1, T_2) \geq k$ .

*Proof:* It suffices to show that one NNI operation changes the number of edges in the shortest path between any two taxa on a tree by at most one. Consider the trees,  $T_1$  and  $T_i$  in Fig. 2, which are reduced to a single, arbitrarily chosen internal edge with two subtrees extending from each of its vertices. The subtrees are denoted by  $S_1, S_2, S_3$  and  $S_4$ . Note that these are subtrees and not necessarily external edges, as indicated by the arrows at the end of each edge. For  $n > 3$ , all unrooted, binary trees with  $n$ -taxa can be reduced to a single internal edge with four subtrees extending from it, as in Fig. 2a and Fig. 2b. Consider the NNI that switches  $S_1$  and  $S_2$  on  $T_1$  without loss of generality. Every possible NNI on  $T_1$  and  $T_i$  will take this form, as an NNI switches two subtrees that are connected by a single internal edge. This results in the tree,  $T_i$ .

We know that the topology of a subtree on  $T_1$  cannot be changed by an NNI about this internal edge. An NNI operation can insert or delete at most one edge between any two taxa. Then, since  $T_i$  is one NNI operation away from  $T_1$ ,  $k - 1 \leq d(T_i, T_2) \leq k + 1$ . Therefore, it takes at least  $k$  NNI operations in order to transform  $T_1$  into  $T_2$ . So, the NNI distance between  $T_1$  and  $T_2$  is at least  $k$ .  $\square$

The lower bound is strict, since we can always find two  $n$ -taxa trees which satisfy  $k$ -IC and which can be rearranged to match in a series of  $k$  NNI moves. For example, in Fig. 3, we have with two  $n$ -taxa caterpillar trees, where  $n = k + 4$ . We can get from  $T_1$  to  $T_2$  in a series of  $k$  NNI operations by simply switching  $i_3$  on  $T_1$  with the pendant edge to the left of it  $k$  times. Since the NNI distance between  $T_1$  and  $T_2$  cannot be less than  $k$ , their NNI distance is  $k$ . Though  $k$ -IC may not provide the best possible lower bound on the NNI distance, it is much quicker to compute than the NNI distance itself.

Using similar methods with  $C_{3x}$  and  $D_{3x}$ , we can show that  $k$ -IC does not determine the Robinson-Foulds

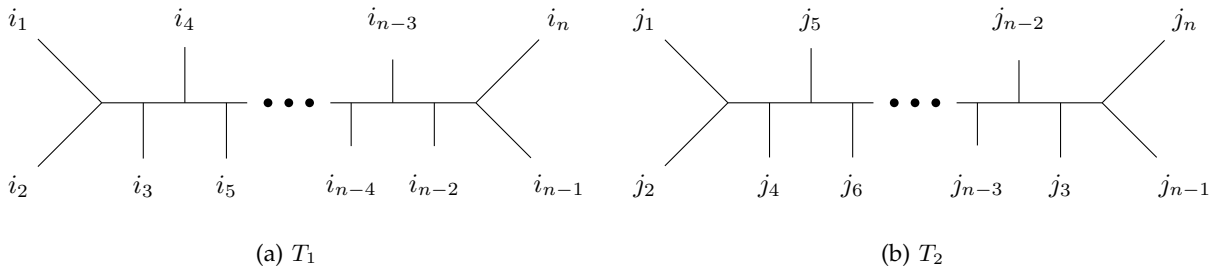


Fig. 3

distance, the path difference distance or the maximum agreement subtree distance. From these comparisons, we see that the  $k$ -IC between two trees describes global similarities between the trees that other common discrete distance metrics do not. This led us to believe that  $k$ -IC is a useful tool for studying pairs of cophylogenetic trees.

One way of describing this global similarity is through the length of the longest path between two taxa on the trees. We use the longest paths on two trees that satisfy  $k$ -IC to quantify the observation that  $k$ -IC describes global similarities between trees. If the difference between the length of the longest paths on two trees is equal to  $x$ , then  $k$  must be at least  $x$ .

**Theorem 4.** *If two  $n$ -taxa trees,  $T_1$  and  $T_2$  precisely satisfy  $k$ -IC, then the absolute value of the difference between the length of the longest path on  $T_1$  and the length of the longest path on  $T_2$  is at most  $k$ . Additionally, if this difference is equal to  $k$ , then taxa on opposite ends of the longest path on  $T_1$  must also be at the ends of the longest simple path on  $T_2$ .*

*Proof:* Without loss of generality, let the longest path on  $T_1$  be greater than the longest path on  $T_2$ . Let  $A, B$  be the taxa which lie at the ends of the longest path on  $T_1$ . In other words,  $\varepsilon_1(A, B) = \max(\varepsilon_1(Q, R))$  for all  $Q, R \in T_1$ . Since  $a, b \in T_2$  can be no farther apart than  $\max(\varepsilon_2(s, t))$ ,  $\varepsilon_1(A, B) - \varepsilon_2(a, b) \geq \max(\varepsilon_1(Q, R)) - \max(\varepsilon_2(s, t))$ . In other words, the difference between the number of edges between  $A$  and  $B$  and the number of edges between  $a$  and  $b$  must be greater than the difference between the longest paths on  $T_1$  and  $T_2$ . So,  $|\max(\varepsilon_1(Q, R)) - \max(\varepsilon_2(s, t))| \leq k$ . If  $|\max(\varepsilon_1(Q, R)) - \max(\varepsilon_2(s, t))| = k$ , then  $a$  and  $b$  must lie along the longest path in  $T_2$ . If they did not,  $\varepsilon_1(A, B) - \varepsilon_2(a, b)$  would be greater than  $k$ , contradicting the fact that  $d(T_1, T_2) = k$ .  $\square$

This result helps to quantify the degree of global congruence between two cophylogenetic trees so that researchers need not rely upon visual inspection.

## 4 THE $(n - 3)$ -INTERVAL NEIGHBORHOOD

The Robinson-Foulds Distance is one of the most commonly used distance metrics in phylogenetics. However, as papers such as [8] mention, the growth of the neighborhood of trees the farthest possible Robinson-Foulds distance away from a given tree can make it difficult to use for analysis of simulated data. In order to analyze these neighborhoods in terms of  $k$ -IC, we define the  $k$ -interval neighborhood. The  $k$ -interval neighborhood ( $IN_k$ ) of a tree,  $T$ , is defined to be the set of all trees which precisely satisfy  $k$ -IC with  $T$ . By analyzing the size of the neighborhood of trees which precisely satisfy the largest possible  $k$ -IC with a given tree, we can gain insight into how  $k$ -IC can be used to analyze simulated data.

### 4.1 Size of the $(n - 3)$ -Interval Neighborhood

**Lemma 1.** *Two  $n$ -taxa trees precisely satisfy at most  $(n - 3)$ -IC.*

*Proof:* Consider two  $n$ -taxa trees,  $T_1$  and  $T_2$ . Then  $T_1$  and  $T_2$  each have exactly  $n - 3$  internal edges [9]. Therefore, two taxa can have at most  $n - 1$  edges between them, and there must be at least 2 edges between them. So, to maximize the  $k$ -interval cospeciation of  $T_1$  and  $T_2$ , let  $\varepsilon_1(I, J) = n - 1$  for some  $I, J \in T_1$  and  $\varepsilon_2(i, j) = 2$  for corresponding  $i, j \in T_2$ . Then  $\varepsilon_1(I, J) - \varepsilon_2(i, j) = n - 3$ . Since this method maximizes  $k$ ,  $d(T_1, T_2) \leq (n - 3)$ .  $\square$

**Corollary 1.** *If two for  $n$ -taxa trees,  $T$  and  $T'$ ,  $d(T, T') = n - 3$ , at least one of them must be a caterpillar tree.*

This follows directly from the fact that, in order for two trees to satisfy  $(n - 3)$ -IC, one of those trees must have a path of length  $n - 1$ . The only trees with such paths are caterpillar trees.

Let  $ub(n)$  denote the number of unrooted binary trees on  $n$  taxa. We know from [10] that

$$ub(n) = (2n - 5)!! = (2n - 5) \times (2n - 3) \times \dots \times 5 \times 3 \times 1.$$

**Theorem 5.** *For an  $n$ -taxa tree,  $T$ , with  $c$  cherries,*

$$|IN_{(n-3)}| = \begin{cases} 2((n - 2)! + 2((n - 4)! - 2(n - 3)!) + 2(2n - 7)!! - (2n - 9)!!) & \text{if } c = 2, \\ c((n - 2)! - 2(n - 4)!(c - 1)) & \text{if } c \geq 3. \end{cases}$$

*Proof:* Case 1:  $c = 2$ . When  $c = 2$ ,  $T$  is a caterpillar tree. So, if  $d(T, T') = n - 3$ ,  $T'$  can have any topology. Suppose  $T$  has two cherries,  $(a, b)$  and  $(c, d)$ . We can build such a  $T'$  from  $T$  either by:

- placing two taxa on opposite cherries of  $T$  into the same cherry on  $T'$ , or
- placing two taxa from the same cherry of  $T$  into opposite ends of  $T'$ , when  $T'$  is a caterpillar.

First, we count the trees obtained by placing two taxa on opposite cherries of  $T$  into the same cherries on  $T'$ . There are four possible cherries to pick from:  $(a, c)$ ,  $(a, d)$ ,  $(b, c)$  and  $(b, d)$ . Since the number of  $n$ -taxa unrooted binary trees with a single cherry fixed is equal to the number of  $(n - 1)$ -taxa unrooted binary trees, fixing these four cherries yields  $4ub(n - 1)$  trees. However, it is possible for cherries  $(a, c)$  and  $(b, d)$  to appear on the same tree

and for cherries  $(a, d)$  and  $(b, c)$  to appear on the same tree. Since the number of  $n$ -taxa unrooted binary trees with two cherries fixed is equal to the number of  $(n-2)$ -taxa unrooted binary trees, there are  $2ub(n-2)$  trees which have been double-counted. So,  $2(2ub(n-1) - ub(n-2)) = 4(2(2n-7)!! - (2n-9)!!)$  is the number of trees which precisely satisfy  $(n-3)$ -IC with  $T$  yielded by putting taxa from opposite cherries in  $T$  into the same cherry in  $T'$ .

The rest of the trees in  $\text{IN}_{(n-3)}(T)$  can be counted by fixing two taxa that are in cherries in  $T$  in opposite cherries of a caterpillar tree  $T'$ . There are  $(n-2)!$  caterpillar trees with  $a$  and  $b$  fixed this way. However,  $(n-3)!$  of those contain cherry  $(a, c)$  and  $(n-3)!$  contain cherry  $(b, d)$ , which we counted in the previous step. Of those  $2(n-3)!$ ,  $(n-4)!$  include both  $(a, c)$  and  $(b, d)$  as cherries, so not all of them are unique. Therefore, there are  $2(n-3)! - (n-4)!$  trees with cherry  $(a, c)$  or  $(b, d)$  which have already been counted. Similarly, there are  $2(n-3)! - (n-4)!$  trees with cherry  $(a, d)$  or  $(b, c)$  which have already been counted. So, by fixing  $a$  and  $b$  in opposite cherries of a caterpillar tree, we yield  $(n-2)! - 2(2(n-3)! - (n-4)!)$  new trees in  $\text{IN}_{(n-3)}(T)$ .

By the same argument, we obtain another  $(n-2)! - 2(2(n-3)! - (n-4)!)$  new trees by fixing  $c$  and  $d$  in opposite cherries with no double counting. Therefore, for a caterpillar tree,

$$\begin{aligned} |\text{IN}_{(n-3)}(T)| &= 2((n-2)! - 2(2(n-3)! - (n-4)!)) + 2(2ub(n-1) - ub(n-2)) \\ &= 2((n-2)! - 2(2(n-3)! - (n-4)! + 2(2(2n-7)!! - (2n-9)!!))). \end{aligned} \quad (5)$$

Case 2:  $c \geq 3$ . When a tree,  $T$ , has  $c$  cherries with  $c \geq 3$ , we know that  $T$  cannot be a caterpillar. Therefore, in order for  $d(T, T') = n-3$  to hold,  $T'$  must be a caterpillar. Additionally, since a caterpillar is the only tree topology with an internal path length of  $n-1$ , we know that for each possible  $T'$ , there must be two taxa that were in a cherry on  $T$  which are in opposite cherries on  $T'$ .

For each cherry,  $(i_l, j_l)$  on  $T$ ,  $1 \leq l \leq c$ , there exist  $(n-2)!$  caterpillar trees with  $i_l$  and  $j_l$  in opposite cherries. For  $i_1$  and  $j_1$ , this amounts to  $(n-2)!$  trees counted in  $\text{IN}_{(n-3)}(T)$ . For  $(i_2, j_2)$ , there are  $(n-2)!$  caterpillar trees with  $i_2$  and  $j_2$  in opposite cherries. However,  $(n-4)!$  have of these have cherries  $(i_1, i_2)$  and  $(j_1, j_2)$  and another  $(n-4)!$  have cherries  $(i_1, j_2)$  and  $(j_1, i_2)$ . So of the  $(n-2)!$  caterpillar trees with  $i_2$  and  $j_2$  in opposite cherries,  $2(n-4)!$  were already counted. In fact, for each  $(i_l, j_l)$ ,  $2(l-1)(n-4)!$  have already been counted for the previous cherries of  $T$ . So

$$\begin{aligned} |\text{IN}_{(n-3)}(T)| &= \sum_{l=1}^c ((n-2)! - 2(l-1)(n-4)!) \\ &= c(n-2)! - 2(n-4)! \sum_{l=2}^c (l-1) \\ &= c(n-2)! - 2(n-4)! \sum_{l=1}^{c-1} l \\ &= c(n-2)! - \frac{2(n-4)!c(c-1)}{2} \\ &= c((n-2)! - (n-4)!(c-1)). \end{aligned} \quad (6)$$

□

## 4.2 Growth of the $(n-3)$ -Interval Neighborhood

The size of the  $\text{IN}_{(n-3)}$  is not constant for every  $n$ -taxa tree,  $T$ . It grows with respect to  $n$  as well as with respect to the number of cherries,  $c$ , on  $T$ . Understanding how  $\text{IN}_{(n-3)}$  grows with respect to tree topology can help us to understand what information  $k$ -interval cospeciation conveys about the two trees we are comparing.

**Theorem 6.** *For all  $n$ -taxa trees with three or more cherries, a tree with three cherries has the smallest  $(n-3)$ -interval neighborhood, and a tree with  $\lfloor \frac{n}{2} \rfloor$  cherries has the largest. The size of the  $(n-3)$ -interval neighborhood increases as a quadratic on  $c$ .*

*Proof:* For  $c \geq 3$ , we have (6). We can expand this formula in order to observe it as a quadratic function on  $c$ .

$$\begin{aligned} |\text{IN}_{(n-3)}| &= c(n-2)! - c(c-1)(n-4)! \\ &= c(n-2)! - c^2(n-4)! + c(n-4)! \\ &= -c^2(n-4)! + c((n-2)! + (n-4)!) \end{aligned} \quad (7)$$

So, as  $c$  increases, the size of  $\text{IN}_{(n-3)}(T)$  is increasing as a concave down quadratic. The vertex of this parabola is located at  $c = \frac{(n-2)! + (n-4)!}{2(n-4)!}$ . So, if  $c$  could ever exceed that value, the size of  $\text{IN}_{(n-3)}(T)$  would start to decrease. In

order to prove that this cannot happen, we show that  $\lfloor \frac{n}{2} \rfloor \leq \frac{(n-2)! + (n-4)!}{2(n-4)!}$  for all relevant values of  $n$ . The equation for the vertex expands to the following quadratic on  $n$ :

$$\begin{aligned} c &= \frac{1}{2} \left( \frac{(n-2)!}{(n-4)!} + 1 \right) \\ &= \frac{1}{2} (n^2 - 5n + 7). \end{aligned} \quad (8)$$

The greatest number of cherries a binary tree can have is  $\lfloor \frac{n}{2} \rfloor$ . So, in order for  $c$  to reach the vertex and for the size  $\text{IN}_{(n-3)}(T)$  to begin decreasing, it must be true that  $\frac{1}{2}(n^2 - 5n + 7) \leq \lfloor \frac{n}{2} \rfloor$ . It suffices to find where  $\frac{1}{2}(n^2 - 5n + 7) \leq \frac{n}{2}$ , which is where

$$\frac{1}{2}n^2 - 3n + \frac{7}{2} \leq 0. \quad (9)$$

The roots of the above quadratic are located at  $n = 3 \pm \sqrt{2}$ . So, the inequality is only true when  $n \leq 3 + \sqrt{2}$ . Therefore, we will consider trees for which  $n = \lfloor 3 + \sqrt{2} \rfloor = 4$ . However, recall that (6) applies only to trees with 3 or more cherries. Any tree with  $n \leq 5$  taxa can only have 2 cherries. So this root of (9) is also not significant.

So, the number of cherries on a given tree,  $T$ , with  $n$  taxa and  $c \geq 3$  will never reach the vertex of (7). Therefore, of all the trees in the set of trees with  $c \geq 3$ , the trees with the largest  $(n-3)$ -IN are those with  $\lfloor \frac{n}{2} \rfloor$  cherries.  $\square$

The following two lemmas will be necessary to compare the size of  $\text{IN}_{(n-3)}(T)$  for trees with  $c \geq 3$  to that of caterpillar trees.

**Lemma 2.** *The inequality,*

$$3ub(n-1) > (n-1)! \quad (10)$$

*holds for all  $n \geq 10$ .*

*Proof:* Base Case:  $n = 10$ . When  $n = 10$ ,  $3ub(n-1) = 3(13)!! = 405405$  and  $(n-1)! = 362880$ . So (10) holds for  $n = 10$ .

Let (10) hold for  $n = x - 1$ . Then  $3ub(n-2) = 3(2x-9)!! > (x-2)!$ . We can expand  $3ub(n-1) = 3(2x-7)!!$  to  $3(2x-7)(2x-9)!!$  and  $(x-1)!$  to  $(x-1)(x-2)!$ . Since  $3(2x-9)!! > (x-2)!$  is true by the induction hypothesis, and  $(2x-7) > (x-1)$  for all  $x > 6$ ,  $3(2x-7)(2x-9)!! > (x-1)(x-2)!$  is true for all relevant  $x$ . Therefore, (10) is true for all  $n \geq 10$ .  $\square$

**Lemma 3.** *For all natural numbers,  $n$ ,*

$$n^3 - 6n^2 + 11n - 6 > \frac{n^3}{2} - 5n^2 + \frac{37n}{2} - 34. \quad (11)$$

*Proof:* Let  $f(n) = n^3 - 6n^2 + 11n - 6$  and  $g(n) = \frac{n^3}{2} - 5n^2 + \frac{37n}{2} - 34$ . The difference,  $f(n) - g(n)$  has two complex roots and one negative root. Observe that  $f(1) = 0$  and  $g(1) = -20$ . Since  $f(n)$  and  $g(n)$  are continuous and never cross in the positive reals, and since we can find a positive value of  $n$  for which  $f(n) > g(n)$ , this inequality holds for all positive  $n$ .  $\square$

**Theorem 7.** *Of all  $n$ -taxa trees, the caterpillar tree has the largest  $(n-3)$ -interval neighborhood.*

*Proof:* We know that (10) is true for  $n \geq 10$ . We note that  $3ub(n-1) = 4ub(n-1) - ub(n-1) < 4ub(n-1) - 2ub(n-2)$  for all  $n \geq 5$ . So we can replace the left hand side of (10) with  $4ub(n-1) - 2ub(n-2)$ . The expression,  $(n-1)!$ , expands to  $(n-1)(n-2)(n-3)(n-4)!$ . Since  $(n-1)(n-2)(n-3)$  expands to  $n^3 - 6n^2 + 11n - 6$  which is greater than  $\frac{n^3}{2} - 5n^2 + \frac{37n}{2} - 34$  for all positive real numbers, we replace the right hand side of 10 with  $(\frac{n^3}{2} - 5n^2 + \frac{37n}{2} - 34)(n-4)!$ . Therefore,

$$4ub(n-1) - 2ub(n-2) > \left( \frac{n^3}{2} - 5n^2 + \frac{37n}{2} - 34 \right) (n-4)! \quad (12)$$

holds. After some algebraic manipulation, we found that (12) is equivalent to the following inequalities:

$$\begin{aligned} 4ub(n-1) - 2ub(n-2) &> \left( \frac{n^3}{2} - \frac{5n^2}{2} + 6 - \frac{n^2}{2} + \frac{n}{2} - 2n^2 + 18n - 40 \right) (n-4)! \\ &> \left( \frac{n}{2}(n^2 - 5n + 6) - \frac{n}{2} \left( \frac{n}{2} - 1 \right) \right) (n-4)! - (2(n^2 - 5n + 6) + 4 - 8n + 24)(n-4)! \end{aligned}$$

When we simplify the above expression, we can see that the following hold as well.

$$\begin{aligned} (2(n^2 - 5n + 6) + 4 - 8n + 24)(n-4)! + 4ub(n-1) - 2ub(n-2) &> \left( \frac{n}{2}(n^2 - 5n + 6) - \frac{n}{2} \left( \frac{n}{2} - 1 \right) \right) (n-4)! \\ 2(n-2)! + 4(n-4)! - 8(n-3)! + 4ub(n-1) - 2ub(n-2) &> \frac{n}{2}(n-2)! - \frac{n}{2} \left( \frac{n}{2} - 1 \right) (n-4)!. \end{aligned} \quad (13)$$

From Theorem 5, we know that the left hand side of this equality is the size of the  $(n-3)$ -interval neighborhood of a caterpillar tree, while the right hand side is that of a tree with  $\lfloor \frac{n}{2} \rfloor$  cherries. Therefore, the  $(n-3)$ -interval neighborhood of a caterpillar tree is larger than that of a tree with  $\lfloor \frac{n}{2} \rfloor$  cherries for all  $n \geq 10$ . We can also compute that this is true for  $n$  between 6 and 9.  $\square$

### 4.3 The $(n-3)$ -Interval Neighborhood as a Proportion of All Trees

In phylogenetics, we use simulations to test how well our methods work on "perfect" data which satisfy all of the assumptions of the given method. For instance, we can use a phylogenetics method to reconstruct a tree, and then compare it to the original tree to see how well the method worked [11]. Distance metrics are useful for conducting these comparisons, but it is important for us to understand exactly what these metrics tell us about the trees we are comparing. So we must understand the growth of the  $(n-3)$ -interval neighborhood as a proportion of all trees in order to understand how  $k$ -IC can help us in analyzing the results of random tree simulations.

In [12], Bryant and Steel characterized the growth of the neighborhoods of trees a given Robison-Foulds Distance away from any given tree. The Robinson-Foulds distance between  $T_1$  and  $T_2$  is defined to be the normalized number of splits induced by  $T_1$  and not by  $T_2$ , and vice versa [13]. Bryant and Steel showed that, for the Robinson-Foulds distance, the proportion of trees that share  $s$  splits follows a Poisson distribution with mean  $\lambda = \frac{c}{2n}$ . For a given tree  $T$ , the expected proportion of trees that share  $s$  splits with  $T$  is given by  $\lambda^s e^{-\lambda/s}$  [12]. When two trees,  $T_1$  and  $T_2$  have  $s = 0$ , this means that they share no splits, and that they are the farthest possible Robinson-Foulds distance apart. The expected value of the proportion of trees which share 0 splits is given by

$$E(s = 0) = e^{-c/2n}. \quad (14)$$

In order to understand how the proportion of trees with  $s = 0$  with  $T$  grows for  $n$  large, we can take the limit of (14) as  $n$  approaches infinity:

$$\lim_{n \rightarrow \infty} e^{-c/2n} = 1.$$

This is true for any fixed  $c$ . So, if we fix a number of cherries on a tree, and let  $n$  approach infinity, it becomes increasingly likely that we will randomly choose a tree that is the farthest possible Robinson-Foulds distance away from  $T$ . However, if we let  $c$  be the maximum number of cherries possible, we have

$$\lim_{n \rightarrow \infty} e^{-\lfloor n/2 \rfloor / 2n} = e^{-1/4} \approx 0.7788.$$

So, the proportion of trees the farthest possible Robinson-Foulds distance away from a given tree approaches 1 as  $n$  approaches infinity for trees with a fixed number of cherries. This proportion also approaches 1 as  $n$  approaches infinity and  $c$  approaches 2. This makes the Robinson-Foulds distance less useful for simulations on random large trees, since almost every resulting tree will be the farthest possible distance away from the comparison tree. This does not tell us all of the possible information about how the phylogenetic method performed.

**Theorem 8.** *The proportion of trees in the  $(n-3)$ -IN of a given tree approaches 0 as  $n$  approaches infinity.*

We used Mathematica 9.0 [14] to compute that

$$\lim_{n \rightarrow \infty} \frac{2((n-2)! + 2((n-4)! - 2(n-3)!) + 2(2n-7)!! - (2n-9)!!)}{(2n-5)!!} = 0$$

and that

$$\lim_{n \rightarrow \infty} \frac{\lfloor n/2 \rfloor ((n-2)! - (\lfloor n/2 \rfloor - 1)(n-4)!)}{(2n-5)!!} = 0.$$

These two cases suffice to show that the proportion of trees in the  $(n-3)$ -interval neighborhood approaches 0 as  $n$  approaches infinity for all trees, since these are the largest two cases. The fact that as  $n$  grows, we become less and less likely to choose a tree that is the farthest distance apart from a given tree may prove more helpful for analyzing simulations on random trees.

## 5 CONCLUSION

When phylogeneticists work to reconstruct the "Tree of Life", it is unlikely that every species relationship will be perfectly accurate. Instead, it is more important that relationships between larger taxonomical units be correct, and that the tree have a high degree of global accuracy. Consider  $C_{3x}$  and  $D_{3x}$  from Fig. 1. For large values of  $x$ , the Robinson-Foulds distance will be quite large, while  $C_{3x}$  and  $D_{3x}$  will still satisfy 2-IC. These large trees would be assigned a worse "score" based on the RF-distance. The fact that  $d(C_{3x}, D_{3x}) = 2$ , however, tells us that these trees are actually very similar to one another. While they are not be a perfect match, we would likely want their

scores to be very good, since their topologies do match, and they exhibit large degree of global congruence. The Robinson-Foulds distance does not facilitate this as well as  $k$ -IC.

If we performed a method on simulated data, and returned a tree with many errors, those errors could all be located at the leaves of a tree with a very similar topology to the original. We may not necessarily want to discard this method, since it did accurately reconstruct the tree topology; however, using the Robinson-Foulds distance may lead to a low "score" on this tree, as it is still almost guaranteed to be the farthest distance away from the original for large trees. In the host-parasite trees from Figure 4 of [2], for example, the trees appear similar through visual comparison, and satisfy 5-interval cospeciation, but their Robinson-Foulds distance is 20. This relatively high Robinson-Foulds distance does not convey how similar the two trees actually are.

We believe that  $k$ -interval cospeciation also has applications to research in the coalescent model of evolution. Papers concerning the coalescent model such as [5] and [15] often discuss the probability of coalescences of lineages and of congruences between gene and species trees. Knowing these probabilities, researchers in this field may be able to determine a likely  $k$  between the gene and species trees in order to help them find the correct trees.

**Open Problem 1.** Given a coalescent model in which  $T$  represents the species tree and  $T_i, T_j$  represent gene trees, find the expected values of  $d(T_i, T_j)$  and  $d(T, T_i)$ .

The field of cophylogenetics is relatively new, and we require new tools to fully understand it. There are still many questions about  $k$ -interval cospeciation to be answered (namely, what can we say about the neighborhood of trees that between 1 and  $(n - 3)$ -IC?). However, our research indicates that  $k$ -interval cospeciation is a useful and unique distance metric. Since  $k$ -interval cospeciation can quantify the degree of global similarity between trees, while allowing for local differences around the leaves, it will prove helpful for those who seek to test and compare cophylogenetic trees.

## ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1358534. Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM12345.

## REFERENCES

- [1] P. Huggins, M. Owen, and R. Yoshida, "First steps toward the geometry of cophylogeny," in *Harmony of Gröbner Bases and the Modern Industrial Society*.
- [2] J. Hughes, M. Kennedy, K. P. Johnson, R. L. Palma, and R. D. Page, "Multiple cophylogenetic analyses reveal frequent cospeciation between peleceniform birds and pectinopygus lice," *Systematic biology*, vol. 56, no. 2, pp. 232–251, 2007.
- [3] G. Refrégier, M. Le Gac, F. Jabbour, A. Widmer, J. A. Shykoff, R. Yockteng, M. E. Hood, and T. Giraud, "Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation," *BMC Evolutionary Biology*, vol. 8, no. 1, p. 100, 2008.
- [4] M. S. Hafner and S. A. Nadler, "Phylogenetic trees support the coevolution of parasites and their hosts," 1988.
- [5] W. P. Maddison, "Gene trees in species trees," *Systematic biology*, vol. 46, no. 3, pp. 523–536, 1997.
- [6] M. Li, J. Tromp, and L. Zhang, "On the nearest neighbour interchange distance between evolutionary trees," *Journal of Theoretical Biology*, vol. 182, no. 4, pp. 463–467, 1996.
- [7] S. Skiena, "Dijkstra's algorithm," *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*, Reading, MA: Addison-Wesley, pp. 225–227, 1990.
- [8] S. Snir and S. Rao, "Quartets maxcut: a divide and conquer quartets algorithm," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 7, no. 4, pp. 704–718, 2010.
- [9] D. F. Robinson, "Comparison of labeled trees with valency three," *Journal of Combinatorial Theory, Series B*, vol. 11, no. 2, pp. 105–119, 1971.
- [10] M. Steel, "Tracing evolutionary links between species," *arXiv preprint arXiv:1402.3771*, 2014.
- [11] J. P. Huelsenbeck, "Performance of phylogenetic methods in simulation," *Systematic biology*, vol. 44, no. 1, pp. 17–48, 1995.
- [12] D. Bryant and M. Steel, "Computing the distribution of a tree metric," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 6, no. 3, pp. 420–426, 2009.
- [13] D. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1, pp. 131–147, 1981.
- [14] W. R. Inc., "Mathematica," 2012.
- [15] N. A. Rosenberg, "The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model," *Evolution*, vol. 57, no. 7, pp. 1465–1477, 2003.